(2)

FINAL REPORT

Contract No.:	N00014-89-J-1393
Period:	1/1/89 - 12/31/90
Date of Submission:	4/15/91
Name of Institution:	San Diego State University
Title of Project:	Content Effects in Mathematics Problem Solving
Principal Investigator:	Sandra P. Marshall

ONR

Summary:

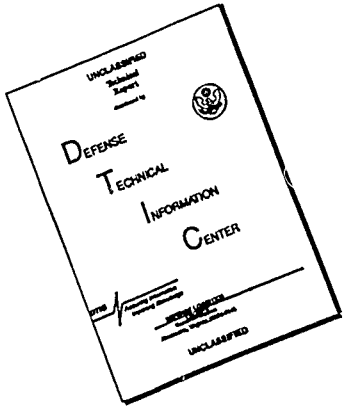
During the two years of this project, we conducted a series of experiments in which we studied two aspects of word problem structure: stereotypy and personal familiarity. The attached paper summarizes our principal findings and conclusions.

Publications and Presentations:

Chipman, Susan F., Marshall, Sandra P., & Scott, Patricia A. Content effects on word problem performance: A possible source of test bias? Conditionally accepted for publication in the *American Educational Research Journal.*

Scott, Patricia A., & Chipman, Susan F. (1990, April). Content effects on word problem performance. In Susan Chipman (Chair), *Penetrating to the mathematical structure of word problems.* Symposium conducted at the Annual Meeting of the American Educational Research Association, Boston.

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Content Effects on Word Problem Performance

A Possible Source of Test Bias?

. Susan F. Chipman

Office of Naval Research

Sandra P. Marshall

Patricia A. Scott

San Diego State University

September, 1990

Revised, March, 1991

Running head: Effects of Word Problem Content

Address correspondence to: Susan F. Chipman, Code 1142CS, Office of Naval Research, 800 N. Quincy Street, Arlington, VA 22217-5000. Email: chipman@nprdc.navy.mil

This research was supported under the Scientific Officer Research Program of the Office of Naval Research via grant N00014-89-J-1393 to the San Diego State University Foundation. The Center for Research in Mathematics and Science Education of SDSU provided facilities for the research. We acknowledge the contribution of John Marshall, who wrote the computer programs to generate the tests and rating instruments.

Abstract

Gender differences in mathematics test performance that favor males are rarely found on tests of computation or other mathematical symbol manipulations. They appear primarily in tests that are labeled as tests of "mathematical reasoning" and consist largely of word problems. The content of word problem cover stories is a possible source of gender bias. Some have suggested that students are discouraged from solving problems for affective reasons when the content of the problem is sex-typed for the opposite sex; cognitive science research on the problem solving processes suggests that familiarity of content would be likely to affect problem solving performance. To test these hypotheses, an experiment was conducted in which underlying mathematics problems were clothed in four different cover stories: masculine, feminine, neutral familiar and neutral unfamiliar. No effect of sex-typing was found; there was an highly significant but small effect of familiarity. Ratings of problem characteristics were also collected, primarily to guide and confirm the realization of the design intentions, and a number of interesting features of the rating results are discussed.

Content Effects on Word Problem Performance: A Possible Source of Test Bias?

Recently there has been much controversy about possible sex bias in the SAT-Math exam, especially as it has been used as the basis of scholarship awards (Rosser, 1989; New York Times, 1989). The SAT-Math scores of young women, on the average, are lower than those of young men, even when one attempts to take account of the small differences in mathematics course background that still exist. In contrast, if one takes course grades as the measure of mathematical performance, women and girls are found to demonstrate equal or better performance (Kimball, 1989) in nearly all studies. A recent analysis by Wainer and Steinberg (1990) showed that when men and women are matched by college mathematics course taken and course grade received, the women are found to have had SAT-Math scores 30-40 points lower. Therefore, the concern about possible sex bias in the SAT is realistic. The most obvious way in which sex bias could arise in the construction of a mathematics test is in the content of the stories used to present mathematics word problems. Indeed, sex differences in mathematics test performance are not as often found as most people believe, and when they are found, it is usually in tests or subtests of mathematics word problems (Chipman & Thomas, 1985; Chipman, 1988; Hyde, Fennema & Lamon, 1990). The diversity and small numbers of items on the Math SAT and similar examinations have led to inconclusive searches for any consistency in the types of items that seem to [illegible]

males or females (Chipman, 1988). For example, one study may report that females do worse on geometry items (Fennema & Carpenter, 1981), another that geometry items favored females or showed no sex difference (Donlon, Hicks & Wallmark, 1980; Becker, 1990). It is, or at least was, true that the word problems on the SAT-Math have tended to have masculine content (Donlon, 1973; Strassberg-Rosenberg & Donlon, 1975). By now, it is evident that post-hoc analyses of item performance will not lead to clearcut conclusions. Direct experimental tests of the hypotheses advanced in the debate over test bias seem to be required.

Experimental studies of the effects of any type of problem content on problem solving performance are surprisingly hard to find. In 1988, both an ERIC search of recent literature and personal inquiries failed to turn up recent studies of problem content effects. Papers by Trible and Higgins and by Barnett, Vos & Sowder in a 1979 ERIC publication edited by Lesh, Mierkiewicz & Kantowski did review a number of relevant studies. Studies of content effects may be rare because mathematics educators have seen content to be rather irrelevant, the goal of the problem solver being to penetrate to the mathematical structure behind the cover story of the problem (Silver, 1979) . However, in a dissertation research project (McCarthy, 1975) a large number of word problems that had been rated for the sex-stereotypy of their content were presented to a large sample of male and female high school students. A reanalysis of these data to [illegible]

effect of sex-typed content demonstrated dramatically large effects (Chipman, 1988). There are also a few very old studies in the literature which provide evidence for such effects of sex-typed content (Milton, 1958) or effects of content familiarity (Brownell & Stretch, 1931). These results suggested that the content of mathematics word problems could be a significant source of test bias and inspired the present study.

There are several hypotheses that one might advance about why such effects could occur. One is the affective hypothesis that students simply shy away from or have difficulty dealing with subject matter that is regarded as the territory of the other sex. (For example, in the study we are reporting here, we did have one male student who refused to solve a problem with a beauty contest cover story, responding to it by writing in, "Let the beauty queen figure it out for herself!") Even if effects of sex-typed content were found, a competing major hypothesis would be that the critical variable is actually familiarity with the content of the items, and that students simply tend to be less familiar with content typed for the opposite sex. Cognitive science studies (as reviewed in Chipman, 1988; Hall, Kibler, Wenger & Truxaw, 1989) of the processes used in solving mathematics word problems have shown that a great deal of thinking is done with the content of the problem, representing it, manipulating it in its own terms, and drawing inferences about the situation that depend on knowledge of the content. Obviously, it is more difficult to draw inferences

about unfamiliar content; even if the need to draw inferences is not at issue, it would be more difficult to maintain representations of unfamiliar content. (As an extreme example, Caldwell & Goldin, 1987, reported that problems verbally stated in terms of abstract mathematical relations are more difficult than those with concrete situations.) Consequently, one would expect poorer performance when problems with unfamiliar content are attempted. At the opposite extreme, very familiar problems may not test the intended problem solving processes at all: Linn & Hyde (1989) report the example of a male student's simply recalling information about sports averages in response to a 1986 SAT item which strongly favored male examinees. In a time-pressed testing situation, omitting items with unfamiliar content or abandoning the attempt to solve them as soon as difficulites arise might be a reasonable strategic decision for an examinee. These hypotheses and the supporting cognitive science literature are discussed at greater length in Chipman (1988).

In order to test these hypotheses, we undertook a study designed to explore the effects of problem content on word problem performance, controlling for the underlying mathematics problem, as had not been done in the McCarthy dissertation.

PRELIMINARY RATING STUDY

In a preliminary study, a large number (96) of mathematics word problems were rated by 50 (25 males; 25 females) San Diego State University students for two characteristics: The sex-stereotypy of the content and the personal familiarity of the content to the individual student doing the rating. The problems came from a variety of sources: 27 from the McCarthy dissertation -- excluding those which really had no story, 50 from old SAT exams -- especially those used by the Princeton Review for practice, 9 from other sources. This preliminary rating study had two primary purposes: to determine whether the variables of sex-stereotypy and personal familiarity could be adequately separated and, secondly, to provide insight into the types of problem content that would receive the types of ratings needed to create the desired experimental design.

Method

Rating Instruments

Fifty different random orders of the 96 problems were prepared. These were used to prepare two rating scale instruments, one of which involved rating sex stereotypy on a 5 point scale from most familiar to males (1) to most familiar to females (5). The other rating scale instrument asked for a rating of the subject's personal familiarity with the content of the problem, again on a 5 point scale from not familiar (1) to very familiar (5). The two

types of ratings were counterbalanced in order in the instrument packages prepared for the subjects.

Subjects and Procedure

Twenty-five male and 25 female San Diego State University students participated in this experiment in order to fulfill a requirement of the introductory psychology course. Thirty minutes were allowed to complete the two sets of ratings, but most subjects finished within 15 minutes. Two experimental sessions were required to obtain the full number of subjects.

Results

The ratings for each problem of sex stereotypy and personal familiarity were averaged over the male and female subjects separately, as well as together. Examination of correlations between the ratings established that the two variables -- sex-stereotypy and personal familiarity -- were somewhat separable, although correlated. Sex stereotypy, as rated by females, had a correlation of .53 with personal familiarity, as rated by females. Sex stereotypy, as rated by males, had a correlation of -.19 with personal familiarity, as rated by males. Males and females agreed strongly in their stereotypy ratings ($r = .94$). The inclusion of 27 items from the McCarthy dissertation made it possible to examine the stability of sex-stereotypy ratings over an elapsed period of about 15 years and a substantial difference in subject populations (New Jersey high school students enrolled in college preparatory mathematics courses vs SDSU college [illegible]

correlation of .96 was found, even though the San Diego State students used a more restricted range of rating values (s.d. of .64 vs. 1.24). Note that the sex-stereotyping of problem content seems to be a notion with a very stable and consistent social consensus. Although it may be difficult to provide any rationale or theory to explain which contents are stereotyped masculine or feminine or considered neutral, it is clear that a solid basis exists for an empirical, operational definition of these categories in our culture.

EFFECTS OF CONTENT ON PROBLEM SOLVING

The ratings obtained from the San Diego State students were used to guide the construction of a set of word problems for use in the experimental study which aimed to examine the effects of both sex-stereotypy and personal familiarity upon problem solving performance. Items which received extreme masculine (less than 2.5) or feminine (greater than 3.5) ratings on the five point scale were selected as starting points. We term the underlying mathematical structure of these problems seeds. Additional problems were written in which the identical mathematical problem was clothed in different cover stories, with the intention of achieving a set of four items for each seed, one of which would be rated as very masculine, one very feminine, one neutral but familiar, one neutral and unfamiliar. Thus, the mathematical structure remains constant under the four cover stories. Table 1 shows one set of four problems developed from the [illegible]

tournament item. The text of all items used, and the ratings these items received are available from the authors. Note that it was the masculinity or femininity of the content of the situation

---

Insert Table 1 about here

---

that was manipulated, not merely the names or sexes of the characters mentioned in problems. Although character names and sexes and personal pronouns have often been used to do analyses of possible bias in test problems (McLarty, Noble, & Huntley, 1988), McCarthy (1975) reported that she did pilot work which indicated that character names and sexes alone did not affect the rated masculinity or femininity of problems. No problem in the set of 96 used in the preliminary rating study was rated as both neutral and unfamiliar so that guidance in writing items that would receive such ratings was not available from the rating data. Often science fiction settings or unspecified exotic cultures were used to evoke ratings of unfamiliarity. The majority of the neutral familiar items involved school settings.

Method

Instrumentation

Problem solving test. Sixteen seeds were used to develop an item bank of 64 items containing matched sets of masculine, feminine, neutral familiar, and neutral unfamiliar items. These items were then used to create a set of mathematics [illegible]

tests with a balanced design in which each subject received a test with 16 problems, one cover story version of each underlying problem, equal numbers (4 each) of masculine, feminine, neutral familiar and neutral unfamiliar items. The result was a set of 32 different tests, with the order of presentation of items randomized within each test. The set of 32 tests yielding a balanced presentation of all 64 problems became the basic unit of the design, to be repeated for sets of male and female subjects. Thus, there were 64 tests, 32 matched pairs for male and female subjects, prepared for the first 64 subjects. To balance for the effects of order of problem presentation, for the second 64 subjects, problems 9-16 of the original test became problems 1-8 of the new version. This entire set of 128 tests was replicated for use with the final 128 subjects. The problems were presented in an open-ended, show-your-work format, rather than as multiple choice exams. This was done in order to avoid the analytic problems posed by guessing responses, including the possibility that the guessing behavior of male and female subjects might differ.

Background questionnaire. A background questionnaire asked whether the subject had successfully completed a mathematics course in each semester of grades 9-12, asked whether the subject had taken calculus in high school, had taken or was currently taking calculus in college, had taken physics in high school, had taken or was currently taking physics or engineering in college, had ever taken a computer programming course.

Problem familiarity ratings. As in the preliminary rating study, subjects were asked to rate the personal familiarity of the problem situation. A seven point scale from not familiar (1) to very familiar (7) was used. All 64 problems were rated. Thirty-two different random orders of presentation of the problems were used.

Sex-stereotypy ratings. Subjects were also asked to rate the sex-stereotypy of the problem situations on a seven point scale from most familiar to males (1) to most familiar to females (7). All 64 problems were rated. Thirty-two different random orders of presentation of the problems were used.

Situation familiarity ratings. This rating instrument contained descriptions of the 64 problem situations that had been stripped of mathematics problem characteristics. Examples are shown in Table 1. Subjects were asked to rate the personal familiarity of the situations described on a seven point scale from not familiar (1) to very familiar (7). Thirty-two different random orders of the situation descriptions were used

Subjects

In the summer and fall of 1989, the first 128 subjects were recruited from among SDSU students. Some were tested during a class period of a summer session introductory psychology course. Because there were too few students to complete the basic design unit of 64, additional paid subjects were recruited. An additional set of 64 subjects was recruited in the fall from the introductory

psychology subject pool. Finally, the second 128 subjects were recruited in the spring of 1990 from the introductory psychology subject pool. Thus, in all there were 256 subjects, 128 females and 128 males. There is every reason to believe that this subject population is reasonably representative of the U.S. college student population at large. In 1988, the mean SAT-Math score of entering freshmen at SDSU was 486 as compared to the national average of 476 in that year, and 96% of accepted applicants took the SAT. Introductory psychology is a very popular course taken by diverse students with many different major fields. With respect to possible self-selection bias, subjects were recruited with an announcement which described a problem solving study, without specifically stating that the problems were mathematical.

Procedures

The procedures differed somewhat for the first group of 128 subjects versus the second group of 128 subjects. Subjects in the first group of 128 were allowed 30 minutes to complete the background questionnaire, which was presented first as a cover page to their test package, and to work the 16 problems; then, they were directed to move on to the rating tasks. These subjects received the problem familiarity and sex-stereotypy rating instruments in counterbalanced order. Subjects in the second group of 128 were given the background questionnaire and situation familiarity instrument at the beginning of the experimental session; when all subjects had completed the ratings, the problem-solving tasks were

distributed. Consequently, the subjects in the second group of 128 had somewhat more time to work on the problems. All experimental sessions for all subjects lasted a total of one hour.

Results

Ratings of Sex-stereotypy and Familiarity

These ratings had been provided by the first 128 subjects only. Ratings for the four classes of problems generally verified the success of the cover story construction. Descriptive statistics are presented in Table 2. The mean stereotypy rating

---

Insert Table 2 about here

---

of the problems intended to be masculine was 3.12, while the mean rating of the problems intended to be feminine was 5.12. The mean rating for the problems intended to be neutral familiar was 4.08, almost precisely in between, while the mean for the problems intended to be neutral and unfamiliar was shifted slightly in the masculine direction, 3.78. The mean familiarity ratings were 4.50 (masculine items), 4.48 (feminine items), 4.48 (neutral familiar items), and 3.83 (neutral unfamiliar items). So, the average familiarity of the masculine, feminine, and neutral familiar problems were the same, while the neutral unfamiliar problems were indeed rated as less familiar than the other types ($t = 7.42$, $p < .001$). As might be expected, masculine problems were rated as somewhat more personally familiar for males (M: 4.67; F: 4.31; ...)

feminine problems as somewhat more personally familiar for females (M: 4.23; F: 4.76). However, as shown in Figures 1a and 1b, the two characteristics do not seem to be seriously confounded, and the

---

Insert Figures 1a, 1b and 1c about here

---

masculine and feminine problems offer a strong contrast on the variable of sex stereotypy. As in the preliminary rating study, the relationships between these variables differed for males and females. Overall sex-stereotypy ratings had a correlation of .42 ($p < .001$) with personal familiarity as rated by females, but an insignificant negative correlation of -.17 with personal familiarity as rated by males. Neutral familiar and neutral unfamiliar items were also quite well contrasted -- see Figure 1c -- but not so cleanly separated as the masculine and feminine problems. Correlations between the ratings collected in the summer and those collected in the fall indicate that sterectypy is a highly reliable rated characteristic ($r = .99$). In addition, as in the preliminary rating study, there was excellent agreement between males and females ($r = .97$). Personal familiarity to the male population or to the female population (combined male and female familiarity ratings, summer vs. fall, $r = .81$) is somewhat less reliable and, of course, not so consistent for males and females ($r = .75$) .

Problem Solving Performance

The first issue to consider is whether the content of the cover stories had an effect on problem solving performance. Analyzing problem solving performance by subjects, we have a classic split-plot design, sex of subject by problem cover story type. Each individual can be regarded as getting a test made up of four subtests, each containing 4 items of a given type. Each subject had random subsets of the problems of each type.

Here we present the analysis for all 256 subjects taken together. Despite the fact that the second set of 128 subjects had a bit more problem solving time available, their performance levels did not differ[1]. Figure 2 plots the mean number correct for males

---

Insert Figure 2 about here

---

and females within each problem cover story type. Performance over the four cover story types was significantly different ($F = 5.03$; $df = 3, 762$; $p < .002$)[2], but the difference in type accounts for little of the total variance. Comparison of SS type to total SS indicates that type accounts for about 2% of the variance. Sheffe post-hoc comparisons among means indicate that neutral unfamiliar cover stories were significantly more difficult than the other three combined ($F = 13.9$; $df = 3, 256$; MS error = .752; $p < .001$).

This analysis also demonstrates that males performed significantly better than females over all types combined ($F = 6.8$; $df = 1, 254$; $p < .012$). There was no significant interaction between gender and cover stories. However, simple main effects comparisons show that males performed significantly better than females on the masculine items ($t = 2.745$, $p < .05$) and marginally better on the neutral familiar items ($t = 2.00$, $.05 < p < .10$). Performance by males and females on the remaining two types of cover stories, the feminine and the neutral unfamiliar, was not significantly different. Again, the effect is small: Comparison of SS gender with SS total indicates that about 2% of the variance is accounted for by gender. Note that the overall level of performance was quite poor, averaging only 31% correct.

Relating Problem Difficulty to Familiarity and Stereotypy

We also explored the relationships among rated familiarity and stereotypy and problem difficulty at the individual item level. For both males ($r = .59$, $p < .001$) and females ($r = .63$, $p < .001$), there was a strong and highly significant correlation between the rated problem familiarity and the p values, the percent correct. In contrast, for neither males ($r = .05$, n.s.) nor females ($r = .17$, n.s.) was the correlation between sex-stereotypy and the p value significant.

Given the magnitude of the familiarity effect reported in the split-plot analysis above, this estimate of the familiarity effect seemed far too large, accounting for 41% of the variance ($F = 4..6$;

df = 1, 62; $p < .0001$)[3]. In this analysis, one no longer enjoys the benefits of the balanced design crossing cover story type with the underlying problem structure. Although subjects were instructed to rate the familiarity of the situation described in the problems and not of the problems themselves, it is likely that the ratings reflected the familiarity of the mathematical structure of the underlying problem. Thus, rated problem familiarity was considered in relation to type of cover story -- masculine, feminine, neutral familiar, and neutral unfamiliar -- and to seed, the underlying mathematics problem, a categorical variable with 16 values. These two variables were independent. Seed accounted for 51% of the variance in rated familiarity and type accounted for 27%. When they were supposed to be rating the familiarity of the problem situation, subjects were responding substantially to the characteristics of the underlying mathematics problem. It was for this reason that situation familiarity ratings were collected from the final 128 subjects.

Correspondingly, an analysis of variance demonstrated that the variance in difficulty accounted for by seed was 89%; adding type increased the variance accounted for to 90%, not a statistically significant increase. All of the predictive power that problem familiarity appeared to have is accounted for by its relation to seed.

Situation Familiarity

The results of the situation familiarity ratings were somewhat surprising. They were much more variable than the problem familiarity ratings, having standard deviations two to three times as large within each of the cover story categories. In the case of "neutral unfamiliar" problems, the situations were rated much less familiar (mean 2.19) than the corresponding problems posed within these situations (mean 3.83). A t-test indicates that this difference is highly significant ($t = 7.87$; $df = 15$, $p < .001$). The correlation between situation familiarity and problem familiarity was highly significant at .54 but still indicates that these ratings measure somewhat different characteristics. The separability of sex stereotyping and familiarity was maintained with this situation familiarity rating. The correlations between stereotypy and situation familiarity were $r = .00$ for males and $r = .36$ ($p < .01$) for females.

Because judgments of problem familiarity seemed to be measuring something closely related to the difficulty of the underlying mathematics problem, it seemed possible that factoring situation familiarity out of problem familiarity might strengthen the relation to difficulty. This was in fact true. The correlation between problem familiarity and p value was .64. The correlation between problem familiarity with situation familiarity factored out and p value was .71. As was true for problem

familiarity, in an analysis of variance, situation familiarity did not provide a significant improvement in variance accounted for over that achieved by the 16 category seed variable.

Omissions

One of the affective hypotheses under consideration was that males and females might selectively omit problems of the opposite sex-type. Omissions were defined as problems on which the test sheet showed no trace of an attempt to solve the problem. Of course, subjects of both sexes were likely to omit problems near the end of the test. More interesting are omissions in which the examinee has made a deliberate choice to omit a problem and go on. Therefore, our analyses consider only omissions prior to the terminal string, omissions followed by at least one attempted problem.

Student performance on the cover story types was examined using a 3 factor log-linear analysis: gender (male, female) by response (correct, incorrect, omit) by cover story type (masculine, feminine, neutral familiar, neutral unfamiliar). The approximately 4000 subject responses were classified under these three dimensions. All possible hierarchical models were considered. The model of best fit includes all three main effects and the two factor interactions of gender by response and response by type ($L = 3.36$; $df = 9$; $p = .95$). The gender by response interaction shows that males were generally more successful than females in solving

the problems and that females were more likely than males to omit problems. There were no differences in the numbers of attempted incorrect responses. The response by type interaction shows that the feminine items were slightly easier than expected (ie. there were more correct responses to these items) and that the neutral unfamiliar items were less likely to be correct and more likely than the others to be omitted.

Looking particularly at the omission data, we observed that females made significantly more omissions than did males ($z = 3.18$, $p < .001$). For both males and females, the masculine, feminine and neutral familiar items were omitted about equally often. Both males and females were more likely to omit the neutral unfamiliar items ($z = 3.52$, $p < .001$). For females, but not for males, the lower number of correct responses on the neutral unfamiliar items seems to have been accounted for entirely by increased omissions.

Mathematics Background, Gender, and Test Performance

Because a statistically significant sex difference in performance was observed, we examined the possibility that sex differences in performance might be accounted for by mathematics course background and/or other mathematics-related course background. The background questionnaire was scored for mathematics course background in the following way: one point was scored for each semester of grades 9-12 in which the subject reported having successfully completed a mathematics course; two points each were added for a yes response to the two items about

having taken calculus in high school or college. Hence the maximum possible score was 12 for a student who had taken a mathematics course in each of the 8 high school semesters and had taken both high school and college calculus. In addition, a variable representing other mathematics-related course background was created by scoring one point each for each yes to the questions about studying physics, engineering and computer programming (3 points maximum).

Complete course background data were available for 232 subjects'. The 118 males had a mean score of 7.37 with an s.d. of 2.52; the 114 females a mean score of 7.60 with an s.d. of 1.89. Obviously, the mathematics backgrounds of the male and female subjects were very similar.

None of the correlations between total test score and mathematics background were significant. Regression analysis confirmed what this suggests. Gender had a statistically significant effect on performance ($p < .02$), in agreement with the result of the split plot analysis reported above, but it was accounting for only 2.4% of the variance in individual performance. Adding the course background variables increased the variance accounted for to only 2.6% and caused the regression to fall below the level of statistical significance.

Discussion

We set out to explore and compare two hypotheses concerning the way in which the content of word problem cover stories might

affect problem solving performance. One was the affective hypothesis that students might avoid or perform poorly on problems with content sex-typed as appropriate for the opposite sex. This hypothesis obtained no support from our results. The other hypothesis had a more cognitive flavor, suggesting that for several reasons one might expect that items with unfamiliar content would result in either omission or poorer problem solving performance. This hypothesis was supported: students of both sexes were more likely to omit problems of neutral but unfamiliar content and less likely to solve such problems correctly. This result was obtained in an experiment in which the underlying mathematical structure of the problems was totally controlled so that we must attribute it to the cover stories. This effect was highly significant statistically, but small in magnitude. There is a possibility that the effect of the familiarity of problem content might become larger under the more stressful conditions of actual testing associated with important consequences, such as the SAT examinations. Omissions, apparent decisions to avoid even attempting a problem, accounted for a large proportion of the effect of familiarity.

Do these results have any message regarding the issue of possible sex bias in word problem examinations? Do they explain sex differences in performance? We did observe a sex difference in problem solving performance in the experiments reported here. Like the content effect, it was very small but highly significant

statistically. The sex difference is not easily explained away by sex differences in mathematics course background. There were none; furthermore, the amount of mathematics background had no effect upon problem solving performance. Obviously, trying to explain tiny effects (the sex difference in problem-solving performance) in terms of other tiny effects (the effect of familiarity of problem content) is a difficult situation to be in. But in fact that situation is characteristic of research concerned with sex differences.

Our speculation that the effect of familiarity might become greater under stress can be accompanied by a speculation that this effect could be greater for females than for males. Above we noted that unfamiliar problems were more often omitted. For females, it appeared that excess omissions totally accounted for the lower level of performance on unfamiliar problems whereas the lower performance by males on such problems was partially attributable to excess omissions and partly attributable to attempted but incorrect solutions. Becker (1990) in a study of the SAT performance of mathematically talented youths noted that females seemed to be more likely to omit problems. The well-documented sex-difference in mathematics anxiety/confidence (Chipman & Wilson, 1985; Hyde, Fennema, Ryan, Frost & Hopp, 1990) may make females more prone to omit problems that appear unfamiliar, especially under the stressful influence of highly important testing situations. On the other hand, we note that there was no sex

difference in performance on the neutral unfamiliar cover stories in the present experiment, even though performance in general was poorer with those problems. It may be that only those students high in mathematical competence can solve problems with such content and that high mathematical competence is equally likely to be found among males or females.

Although we were not able to demonstrate it in this study, it may well be that differences in the familiarity of item content, most likely along with gender differences in confidence and responses to uncertainty and time pressure, are substantially responsible for the often observed gender differences in word problem performance. Potentially, the validity of our speculations about the effects of stressful genuine testing situations could be tested by those who are in a position to manipulate the content of actual tests like the SAT examinations. Probably, manipulating the familiarity of neutral content items would be politically acceptable, partially because it would also be interesting as another approach to measuring mathematical competence. Problem solving performance that is resistant to disruption by unfamiliar content is certainly a goal of mathematics education.

Findings from the Rating Experiments

A number of interesting findings arose in the experiments which obtained ratings of problem familiarity and sex-stereotyping. The most striking of these is the fact that a very rapid rating of familiarity provides a good index of problem difficulty. These

ratings required only about 5 seconds per item! Although the subjects were instructed to rate the familiarity of the situation described in the problem's cover story, our analyses indicated that the ratings were primarily responsive to the characteristics of the underlying mathematics problem, common to all four cover story versions. It is an interesting cognitive question how such a judgment can be made so rapidly. The result concerning the relationship between the rapid ratings and actual problem-solving performance does not necessarily imply that the subjects are analyzing the mathematical structure of the problem in so brief a time. Other research directed at determining whether students can notice and make use of the similarity of structure of successively presented mathematics problems (Reed, 1987, 1988) suggests that they are not performing such an analysis. It could be that students recognize and respond to stereotypic problem gestalts, like river current problems (cf. Hinsley, Hayes, & Simon, 1977; Mayer, 1981), but the process of writing different cover stories made the problems much less stereotyped. A plausible hypothesis is that the complex relationships involved in a difficult multi-step word problem are reflected in the syntactic complexity of the language required to express them. Determining the nature of this judgement process would require further research.

The result does suggest the possibility that ratings of problem familiarity might be an effective, low-cost method by which test developers could do preliminary screening of items for

difficulty. There is also a possibility that this technique could be used to screen for items that might be prone to display large sex or other group differences in problem solving performance. As a very preliminary check of this possibility, we examined the familiarity ratings of the sports average item which Loewen, Rosser & Katzman (1988) reported to have shown a very large sex difference in performance. Would the rating data have spotted the problematic nature of this item? It seems so. This item received a personal familiarity rating of 5.08 from males and 4.42 from females. This difference of .66 is huge in relation to a mean difference of .07. (This item also received one of the most extreme ratings for masculine stereotypy, along with a couple of other sports-related items.) Items which prove to show very large gender differences seem to be idiosyncratic and difficult for test constructors to identify a priori; it would be valuable to have a quick screening method to eliminate them. We leave the further investigation of the potential of this screening method to those with a specific interest in its practical application.

Another interesting result in the familiarity ratings was the great difference in variance of the two types of familiarity ratings. In retrospect, we believe that the much smaller variability of the problem familiarity ratings is due to the fact that all items had in common the familiar component of being confronted with a mathematics word problem. This result and our

interpretation of it are reminiscent of Tversky's (1977) theory about the determination of similarity ratings.

Stereotypy ratings. These ratings had no real predictive power for problem-solving performance. Yet, it is interesting that they show a rock-solid social consensus between males and females and between different groups of subjects. It is interesting that they show no change between the McCarthy study and the present one -- across a 15-year period that most believe to have shown radical change in the social roles of men and women. The presence of a correlation with personal familiarity for females, but not for males, is also interesting, although it seems to have a simple explanation in the fact that stereotypically feminine content tends to involve household activities that are also very familiar to males. In addition, one should note that for both males and females, it is not necessarily the case that they feel personally highly familiar with things that are typed for their sex. This fact can be observed in Figure 1.

Implications for Item Construction and Selection

Because of the strong social consensus associated with it, sex stereotyping of problem content should be very easy for test developers to avoid. Nevertheless, in the past, it has not been avoided, demonstrating a lack of concern for this possible inequity. Personal familiarity is the variable that really seems to matter in performance. However, the asymmetrical structure of the correlations between sex-typing and personal familiarity, [illegible]

males and females suggests that the use of masculine content may put female examinees at a disadvantage with respect to content familiarity that has not been and could not be compensated for by a balancing use of feminine content. We suggest therefore that it may be advisable to use sex-neutral content (or neutral plus feminine content, but no doubt that option would raise objections). Unfortunately, sex differences in personal familiarity -- the variable that really counts -- are much more complex and more difficult to predict intuitively. Indeed we note that both our stereotypy rating instrument and that of McCarthy (1975) actually asked subjects to rate relative familiarity to males and females, but the results of those ratings were shown to be quite different from the results of ratings of personal familiarity made by males and females. Therefore, the additional step of screening items for their familiarity to all groups of interest, minority groups as well as males and females, seems warranted. As we note above, this process would also enable systematic variation of content familiarity, an assessment approach of some interest for its own sake. Given the rapidity with which the relevant judgements can be made, it is feasible to screen items in this way.

No Effect of Mathematics Courses

Finally, we must comment on the rather surprising result that the amount of mathematics course background that the student had did not affect problem-solving performance. Commonly, mathematics course background is found to affect performance on mathematics

tests (Fennema & Sherman, 1977; Wise, 1985), but usually word problems are only a small fraction of the items tested. Our results seem to lend support to the call, made by the National Council of Teachers of Mathematics and others, for increasing emphasis on instruction in problem solving. The absolute level of performance that we found was poor. One could argue that our sample of problems was particularly difficult. However, the fact that the history of mathematics instruction had no value in predicting problem solving performance seems a more disturbing commentary on what is going on in the mathematics instruction that these students received. Against the background of generally poor performance, and apparently ineffective instruction, the devotion of intense energy to interpreting or explaining small sex differences in performance seems a misplaced priority. Perhaps it would be advisable to take a more direct approach to such issues as equity in the award of scholarship opportunities and to turn these energies to improving instruction in problem solving for all students.

REFERENCES

Barnett, J., Vos, K. & Sowder, L. (1979). A review of selected literature in applied problem solving research. In: R. Lesh, D. Mierkiewicz & M. Kantowski (Eds) Applied Mathematical Problem Solving (pp. 73-110). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education, The Ohio State University.

Becker, B.J. (1990) Item characteristics and gender differences on the SAT-M for mathematically able youths. American Educational Research Journal, 27, 65-88.

Brownell, W.A. & Stretch, L.B. (1931). The effect of unfamiliar settings on problem solving. Durham, NC: Duke University Press.

Caldwell, J.H. & Goldin, G.A. (1987). Variables affecting word problem difficulty in secondary school mathematics. J. Research in Mathematics Education, 18, 187-196.

Chipman, S.F. (1988). Word problems: Where test bias creeps in. Paper presented at the annual convention of the American Educational Research Association, New Orleans, April, 1988. (ERIC Document Reproduction Service No. TM 012 411)

Chipman, S.F. & Thomas, V.G. (1985). Women's participation in

mathematics: Outlining the problem. In S.F. Chipman, L.R. Brush & D.M. Wilson (Eds) Women and mathematics: Balancing the equation (pp. 1-24). Hillsdale, NJ: Erlbaum.

Chipman, S.F. & Wilson, D.M. (1985). Understanding mathematics course enrollment and mathematics achievement: A synthesis of the research. In S.F. Chipman, L.R. Brush & D. Wilson (Eds) Women and mathematics: Balancing the equation (pp. 275-328). Hillsdale, NJ: Erlbaum.

Donlon, T.F. (1973). Content factors in sex differences on test questions (ETS RM-73-26). Princeton, NJ: Educational Testing Service.

Donlon, T.F., Hicks, M.M., & Walmark, M.M. (1980). Sex differences in item responses on the Graduate Record Examination. Applied Psychological Measurement, 4, 9-20.

Fennema, E., & Carpenter, T.P. (1981) Sex-related differences in mathematics: Results from the National Assessment. Mathematics Teacher, 74, 554-559.

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization, and affective factors. American Educational Research Journal, 14, 51-71.

Hall, R., Kibler, D., Wenger, E. & Truxow, C. (1989) Exploring the episodic structure of algebra story problem-solving. Cognition and Instruction, 6, 223-283.

Hinsley, D.A., Hayes, J.R. & Simon, H.A. (1977). From words to equations: Meaning and representation in algebra word problems. In M.A. Just & P.A. Carpenter (Eds), Cognitive processes in comprehension, (pp. 89-106). Hillsdale, NJ: Erlbaum.

Hyde, J.S., Fennema, E. & Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107, 139-155.

Hyde, J.S., Fennema, E., Ryan, M., Frost, L.A. & Hopp, C. (1990). Gender difference in mathematics attitudes and affect: A meta-analysis. Psychology of Women Quarterly, 14, 299-324.

Kimball, M.M. (1989). A new perspective on women's math achievement. Psychological Bulletin, 105, 198-214.

Linn, M.C. & Hyde, J.S. (1989). Gender, Mathematics, and Science. Educational Researcher, 18, 17-19,22-27.

Loewen, J.W., Rosser, P. & Katzman, J. (1988). Gender bias in SAT items. Paper presented at the annual convention of the American

Educational Research Association, New Orleans, April, 1988.

Mayer, R.E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories and templates. Instructional Science, 10, 135-175.

McCarthy, K.A. (1976). Sex bias in tests of mathematical aptitude. (Doctoral dissertation, City University of New York). (University Microfilms, 76-11, 629)

McLarty, J.R., Noble, C., & Huntley, R. (1988). Effects of item wording on sex bias. Paper presented to the annual meeting of the National Council on Measurement in Education, New Orleans, April, 1988.

Milton, G.A. (1958). Five studies of the relation between sex role identification and achievement in problem solving. Technical Report No. 3, Department of Industrial Engineering and Department of Psychology, Yale University, New Haven, CT.

New York Times, Education Supplement (August 6, 1989).

Reed, S.K. (1987). A structure-mapping model for word problems. Journal of Experimental Psychology: Learning, Memory and Cognition, 13, 124-139.

Reed, S.K. (1988). Schema-based theories of problem solving. (Technical Report). San Diego: San Diego State University, Center for Research in Mathematics and Science Education.

Rosser, P. (1989). The SAT gender gap: Identifying the causes. Washington, DC: Center for Women Policy Studies.

Silver, E.A. (1979). Student perception of relatedness among mathematical verbal problems. Journal for Research in Mathematics Education, 10, 195-210.

Strassberg-Rosenberg, B. & Donlon, T.P. (1975). Content influences on sex differences in performance on aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education. (ERIC Document Reproduction Service No. ED 110-493; TM 004 766)

Trimble, H.C. & Higgins, J.L. (1979). Problems, applications, interest, and motivation. In: R. Lesh, D. Mierkiewicz & M. Kantowski (Eds) Applied Mathematical Problem Solving, (pp. 25-35). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education, The Ohio State University.

Tversky, A. (1977) Features of similarity. Psychological Review, 84, 327-352.

Wainer, H. & Steinberg, L.S. (1990) Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. Manuscript submitted for publication.

Wise, L.L. (1985). Project TALENT: Mathematics course participation in the 1960s and its career consequences. In S. F. Chipman, L. R. Brush, & D. M. Wilson (Eds) Women and mathematics: Balancing the equation, (pp. 25-58). Hillsdale, NJ: Erlbaum.

Footnotes

1. The correlation between the p values for the first 128 subjects and those for the second 128 was .92 and the absolute numbers of items correct were nearly identical also. The overall (males and females combined) performance was 31% correct in the first group and 33% correct in the second, 34% and 36% for males, 27% and 31% for females. The outcome of an analysis of the first 128 subjects only was the same, showing the same significant effects, though of course the level of statistical significance is greater with increased N.

2. This difference remains highly significant even under the conservative Greenhouser-Geisser correction.)

3. Some might be concerned that the subjects in the first group of 128 had actually just seen 16 of the 64 problems immediately prior to the request to judge the personal familiarity of the problem situations, but the correlation between the personal familiarity as judged by the first 128 subjects and the p values (percent correct) as estimated for the first 128 only and the correlation of personal familiarity as judged by the first group and p values estimated by the second group were identical, .63.

4. Due to experimental error, the first few subjects did not receive the background questionnaire.

TABLE 1

PROBLEMS

Masculine Version:
A hockey team won 3/4 of its games, lost 1/5 of them and tied the rest. If the hockey team tied 10 games, how many did it play altogether?

Feminine Version:
A rising young beauty queen won 3/4 of the local contests that she entered, was out of the running in 1/5 of them and was runner-up in the rest. If she was runner-up in 10 contests, how many did she enter altogether?

Neutral-Familiar Version:
Central High School's "It's Academic" team won 3/4 of the competitions that they entered, lost 1/5 of them and tied the rest. If they tied 10 games, how many did they play altogether?

Neutral-Unfamiliar Version:
The new vaccine protected 3/4 of the test animals from catching the disease to which they were exposed, but 1/5 of the animals died, and the rest became very ill. If 10 animals became very ill, how many animals were there altogether in the test?

SITUATION DESCRIPTIONS

Masculine Version:
A hockey team winning, losing and tying games.

Feminine Version:
A rising young beauty queen winning, losing and being a runner-up in local beauty contests.

Neutral-Familiar Version:
Central High School's "It's Academic" team winning, losing and tying competitions.

Neutral-Unfamiliar Version:
A new vaccine protecting some test animals from catching a disease, while others die or become very ill.

Table 2
Summary of Item Data

			COVER STORY TYPE			
			Masculine	Feminine	Neutral Familiar	Neutral Unfamiliar
PROBLEM FAMILIARITY RATING:						
	Males	X	4.67	4.23	4.38	3.89
		SD	.64	.51	.51	.37
	Females	X	4.31	4.76	4.60	3.78
		SD	.54	.50	.53	.36
SITUATION FAMILIARITY RATING:						
	Males	X	4.24	3.76	3.81	2.34
		SD	1.80	1.00	1.43	.83
	Females	X	3.77	4.71	4.04	2.04
		SD	1.99	1.08	1.61	.96
STEREOTYPE RATING:						
	Males	X	3.09	5.03	4.08	3.78
		SD	.33	.23	.27	.22
	Females	X	3.17	5.18	4.08	3.79
		SD	.34	.26	.29	.28
P VALUES:						
	Males	X	.37	.37	.37	.29
		SD	.23	.24	.25	.22
	Females	X	.27	.31	.31	.26
		SD	.21	.22	.25	.21

Figure Captions

Figure 1. Figure 1a plots the sex-stereotypy ratings versus the personal familiarity ratings of the groups of problems intended to be masculine and feminine, as judged by females subjects only. Figure 1b is the same, but for male subjects only. Figure 1c plots the sex-stereotypy and personal familiarity ratings for the two groups of problems intended to be neutral-familiar and neutral-unfamiliar, as judged by male and female subjects combined.

Figure 2. Problem solving performance. The number of problems correct (maximum 4) is plotted for each of the cover story types, designated as subtests, for male and female subjects separately.

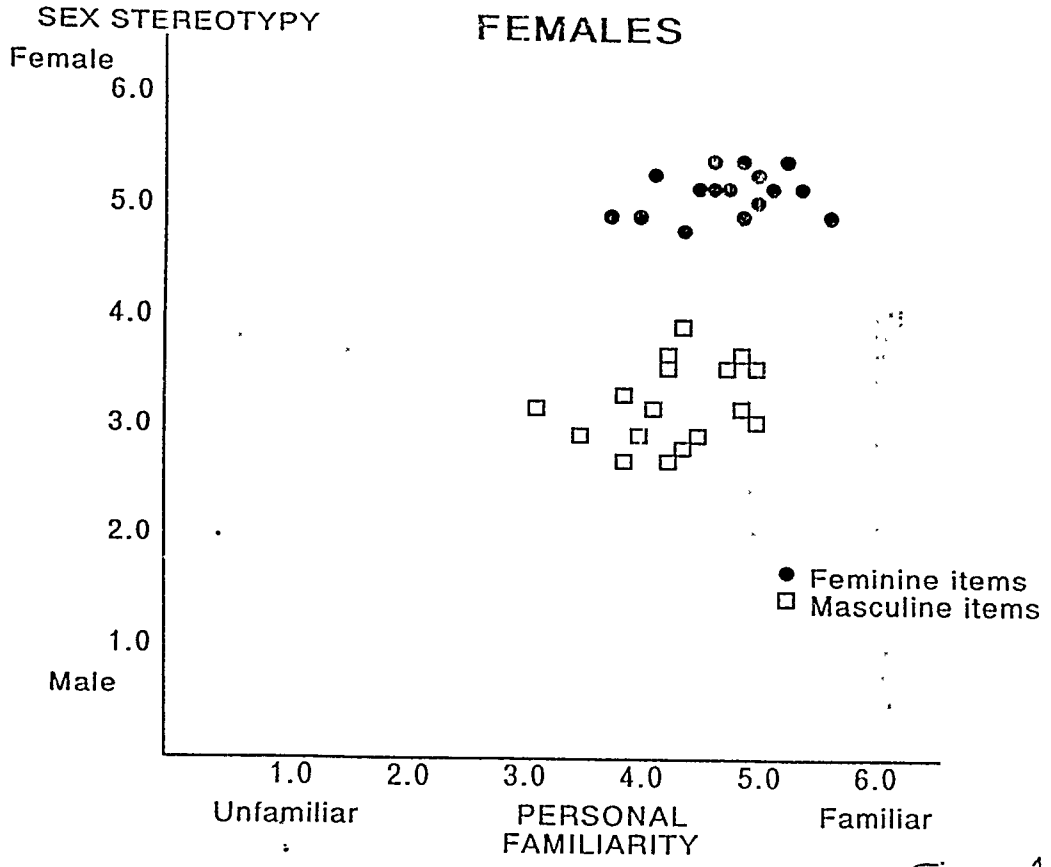

SEX STEREOTYPY
FEMALES
Female
6.0
5.0
4.0
3.0
2.0
1.0
Male
Feminine items
Masculine items
1.0
2.0
3.0
4.0
5.0
6.0
Unfamiliar
PERSONAL FAMILIARITY
Familiar

Figure 1a

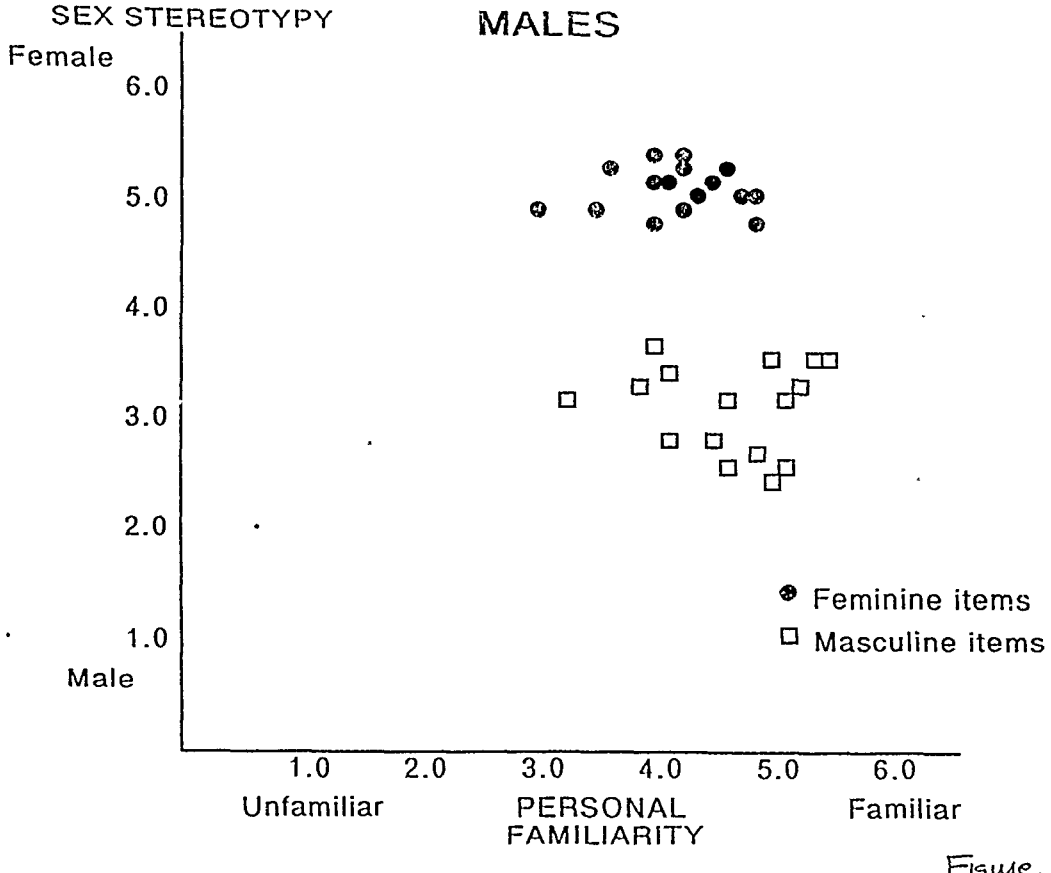

SEX STEREOTYPY
MALES
Female
6.0
5.0
4.0
3.0
2.0
1.0
Male
Feminine items
Masculine items
1.0
2.0
3.0
4.0
5.0
6.0
Unfamiliar
PERSONAL FAMILIARITY
Familiar

Figure 1b

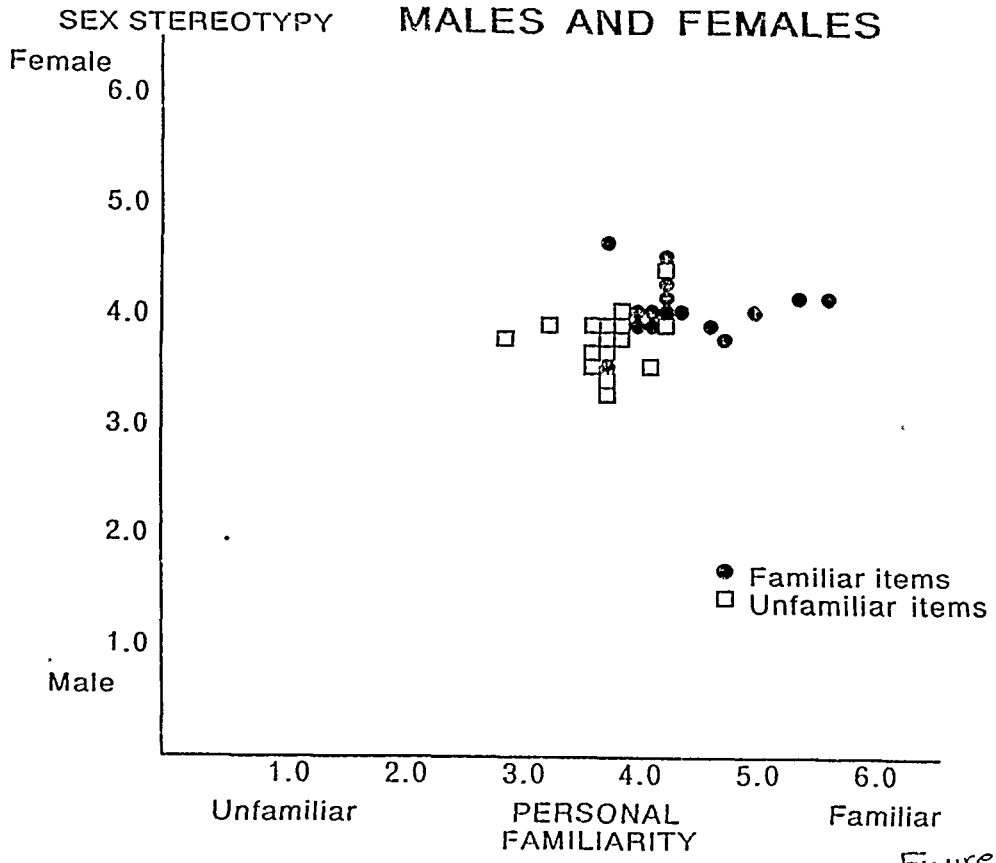

SEX STEREOTYPY
MALES AND FEMALES
Female
6.0
5.0
4.0
3.0
2.0
1.0
Male
Familiar items
Unfamiliar items
1.0
2.0
3.0
4.0
5.0
6.0
Unfamiliar
PERSONAL FAMILIARITY
Familiar

Figure 1c

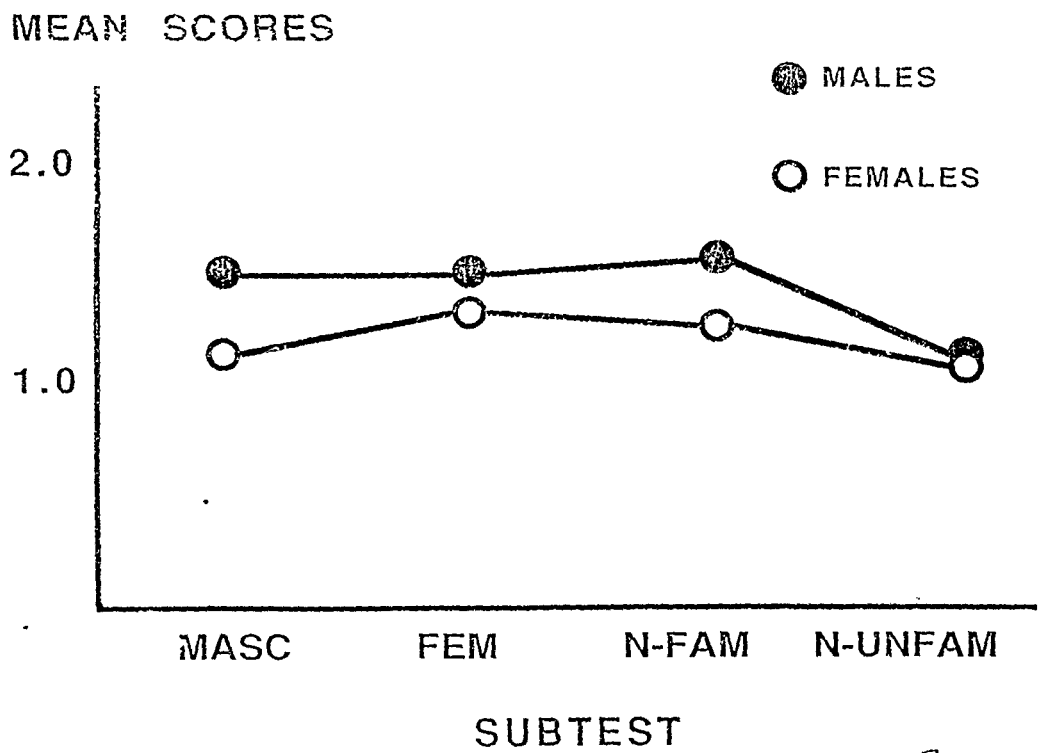

MEAN SCORES
MALES
FEMALES
2.0
1.0
MASC
FEM
N-FAM
N-UNFAM
SUBTEST

Figure 2